Introduction for "An Introduction to Wavelet Analysis"

In this chapter, we give an overview on multiresolution analysis, wavelet series and wavelet estimators in the classical setting. By `classical' or `first-generation' wavelets, we mean wavelets that were constructed initially to analyze signals observed at equispaced design points and having a sample size which is a power of two. The `second-generation' wavelet basis presented in the subsequent chapters will release these two constraints.

If one wants to analyze a function of time with a series expansion, the first idea that comes probably into one's mind is to use a Fourier series, i.e. decompose the function into sine and cosine at different frequencies. In this process, we hope that only a few coefficients in the series will carry most of the information about the signal. Certain smooth functions have such an `economical' Fourier expansion. However, for most functions, a good Fourier series approximation requires numerous sine and cosine basis functions. Indeed, the sine functions have a precise frequency but are not localized in time, hence a localized information in the signal like a discontinuity will affect all the coefficients of the series. This drawback lead the researchers to look for more efficient bases, that is, bases which are localized both in time and in frequency. We will see in Multiresolution analysis and wavelets that a wavelet basis offers this property.

This chapter is structured as follows. We begin by recalling some notations in Function spaces: notion and notations. Next Multiresolution analysis and wavelets introduces the multiresolution analysis, the wavelet functions, and gives some simple examples of wavelet bases. Fast wavelet transform explains how to decompose a signal using the wavelet transform. Such wavelet transforms, also called `decimated', lack the property of translation-invariance. Non-decimated wavelet transform presents a widely used trick to make a wavelet transform translation-invariant. Since the main goal of a wavelet series is to provide a good approximation of a function belonging to a given space, Approximation of Functions introduces some fundamental notions to measure the quality of such approximation. Finally, Nonparametric regression with wavelets presents how to construct a nonparametric regression estimator using wavelets. First, the classical case of equally spaced design is considered. In the last part of Nonparametric regression with wavelets, we review some existing methods that deal with irregular designs.

Function spaces: notion and notations

A Hilbert space is a complete normed space whose norm is indexed by an inner (or scalar) product.

Two disjoint subspaces A and B of a space S form a direct sum decomposition of S if every element of S can be written uniquely as a sum of an element of A and an element of B. The notation $S = A \oplus B$ is then used.

A measurable function f belongs to the Lebesgue space $L_p(\mathbb{R})$, $1 \leq p < \infty$ if
Equation:

$$\|f\|_p = \left(\int_{-\infty}^{+\infty} |f(x)|^p \right)^{1/p} < \infty \ .$$

An example of a Hilbert space is the Lebesgue space $L_2(\mathbb{R})$ of measurable and square integrable functions. Indeed, the norm $\|\cdot\|_2$ is induced by the scalar product
Equation:

$$\langle f, g \rangle = \int f(x) g(x) dx \ ,$$

where $g(x)$ denotes the complex conjugate of $g(x)$. Two functions are said to be orthogonal in $L_2(\mathbb{R})$ if their inner product is zero.

The Lebesgue measure can be replaced by a more general measure μ, leading to the weighted space $\mathbf{L}_2(\mu)$, which has as inner product
Equation:

$$\langle f, g \rangle_{d\mu} = \int f(x) g(x) d\mu(x)$$

and which contains the functions that have a finite norm $\|f\|_{d\mu} := \sqrt{\langle f\,, f\rangle}_{d\mu} < \infty$.

A countable subset $\{f_k\}$ of functions belonging to a Hilbert space is a Riesz basis if every element f of the space can be written uniquely as $f = \sum_k c_k f_k$, and if positive constants A and B exist such that
Equation:

$$A\|f\|_2^2 \leq \sum_k |c_k|^2 \leq B\|f\|_2^2 \ .$$

A Riesz basis is an orthogonal basis if the f_k are mutually orthogonal. In this case, $A = B = 1$.

Multiresolution analysis and wavelets

Definition of subspaces and of scaling functions

A natural way to introduce wavelets is through the multiresolution analysis. Given a function $f\in L_2\left(\mathbb{R}\right)$, a multiresolution of $L_2\left(\mathbb{R}\right)$ will provide us with a sequence of spaces $V_j, V_{j+1}, ...$ such that the projections of f onto these spaces give finer and finer approximations (as $j\to\infty$) of the function f.

Note: (Multiresolution analysis (MRA) in the first generation) A **multiresolution analysis** of $L_2\left(\mathbb{R}\right)$ is defined as a sequence of closed subspaces $V_j\subset L_2\left(\mathbb{R}\right), j\in\mathbb{Z}$ with the following properties:

1. **Equation:**

$$...\subset V_{-1}\subset V_0\subset V_1\subset ...$$

2. The spaces V_j satisfy
   Equation:

$$\bigcup_{j\in\mathbb{Z}} V_j \text{ is dense in } L_2\left(\mathbb{R}\right) \text{ and } \bigcap_{j\in\mathbb{Z}} V_j=\left\{0\right\}\ .$$

3. If $f\left(x\right)\in V_0, f\left(2^jx\right)\in V_j$, i.e. the spaces V_j are scaled versions of the central space V_0.
4. If $f\in V_0, f\left(.-k\right)\in V_0, k\in\mathbb{Z}$, that is, V_0 (and hence all the V_j) is invariant under translation.
5. There exists $\phi\in V_0$ such that $\left\{\phi(x-k); k\in\mathbb{Z}\right\}$ is a Riesz basis in V_0.

We will call `level' of a MRA one of the subspaces V_j. From [link], it follows that, for fixed j, the set $\left\{\phi_{jk}\left(x\right)=2^{j/2}\phi\left(2^jx-k\right); k\in\mathbb{Z}\right\}$ of scaled and translated versions of ϕ is a Riesz basis for V_j. Since $\phi\in V_0\subset V_1$, we can express ϕ as a linear combination of $\left\{\phi_{1,k}\right\}$:
Equation:

$$\phi\left(x\right)=\sum_{k\in\mathbb{Z}} h_k\phi_{1,k}\left(x\right)=\sqrt{2}\sum_{k\in\mathbb{Z}} h_k\phi\left(2x-k\right)\ .$$

[link] is called the **two-scale equation** or **refinement equation**. It is a fundamental equation in MRA since it tells us how to go from a **fine level**V_1 to a **coarser level**V_0. The function ϕ is called the **scaling function**.

As said before, the spaces V_j will be used to approximate general functions. This will be done by defining appropriate projections onto these spaces. Since the union of all the V_j is dense in $L_2\left(\mathbb{R}\right)$, we are guaranteed that any given function of $L_2\left(\mathbb{R}\right)$ can be approximated arbitrarily close by such projections. As an example, define the space V_j as
Equation:

$$V_j=\left\{f\in L_2\left(\mathbb{R}\right); \forall k\in\mathbb{Z}, f|_{\left[2^{-j}k, 2^{-j}(k+1)\right[}=\text{constant}\right\}$$

Then the scaling function $\phi\left(x\right)=1_{[0,1)}\left(x\right)$, called the Haar scaling function, generates by translation and dilatation a MRA for the sequence of spaces $\left\{V_j, j\in\mathbb{Z}\right\}$ defined in [link], see [link], [link].

The detail space and the wavelet function

Rather than considering all the nested spaces V_j, it would be more efficient to code only the information needed to go from V_j to V_{j+1}. Hence we consider the space W_j which complements V_j in V_{j+1} :

Equation:

$$V_{j+1} = V_j \oplus W_j \ .$$

The space W_j is not necessarily orthogonal to V_j, but it always contains the **detail** information needed to go from an approximation at resolution j to an approximation at resolution $j+1$. Consequently, by using recursively the equation [link], we have for any $j_0 \in \mathbb{Z}$, the decomposition

Equation:

$$L_2\left(\mathbb{R}\right) = V_{j_0} \oplus \oplus_{j=j_0}^{\infty} W_j \ .$$

With the notational convention that $W_{j_0-1} := V_{j_0}$, we call the sequence

$\left\{W_j\right\}_{j \geq j_0-1}$ a **multiscale decomposition** (**MSD**).

We call ψ a wavelet function whenever the set $\{\psi(x-k);\ k \in \mathbb{Z}\}$ is a Riesz basis of W_0. Since $W_0 \subset V_1$, there also exist a refinement equation for ψ, similarly to [link]:

Equation:

$$\psi\left(x\right) = \sqrt{2} \sum_k g_k \phi\left(2x-k\right) \ .$$

The collection of wavelet functions $\left\{\psi_{jk} = 2^{j/2}\psi\left(2^j x - k\right); k \in \mathbb{Z}, j \in \mathbb{Z}\right\}$ is then a Riesz basis for $L_2\left(\mathbb{R}\right)$. One of the main features of the wavelet functions is that they possess a certain number of vanishing moments.

Note:A wavelet function $\psi(x)$ has N**vanishing moments** if $\int \psi\left(x\right) x^p dx = 0,\ p = 0, ..., N-1$.

We now mention two interesting cases of wavelet bases.

Orthogonal bases

In an **orthogonal multiresolution analysis**, the spaces W_j are defined as the orthogonal complement of V_j in V_{j+1}. The following theorem tells us one of the main advantages of such a MRA.

Note:([link], Theorem 5.1.1) If a sequence of closed subspaces $\left(V_j\right)_{j \in \mathbb{Z}}$ in $L_2\left(\mathbb{R}\right)$ satisfies [link], and if, in addition, $\{\phi(x-k), k \in \mathbb{Z}\}$ is an orthogonal basis for V_0, then there exists one function $\psi(x)$ such that $\{\psi(x-k);\ k \in \mathbb{Z}\}$ forms an orthogonal basis for the orthogonal complement W_0 of V_0 in V_1.

An immediate consequence of [link] is that $\{\psi_{jk}, k \in \mathbb{Z}\}$ constitutes an orthogonal basis for the orthogonal complement W_j of V_j in V_{j+1}. In this section, let $\mathcal{P}_j$ (resp. $\mathcal{Q}_j$) be the **orthogonal** projection operator onto V_j (resp. W_j). The orthogonal expansion

Equation:

$$\begin{aligned} f &= \mathscr{P}_{j_0} f + \sum_{j=j_0}^{\infty} \mathscr{Q}_j f \\ &= \sum_k \langle f, \phi_{j_0,k} \rangle \phi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_k \langle f, \psi_{jk} \rangle \psi_{jk} \end{aligned}$$

tells us that a first, coarse approximation of f in V_{j_0} is further refined with the projection of f onto the detail spaces W_j.

[link] shows two examples of orthogonal wavelet functions. The first is the Haar wavelet, associated to the Haar scaling function defined in "Definition of subspaces V j and of scaling functions".

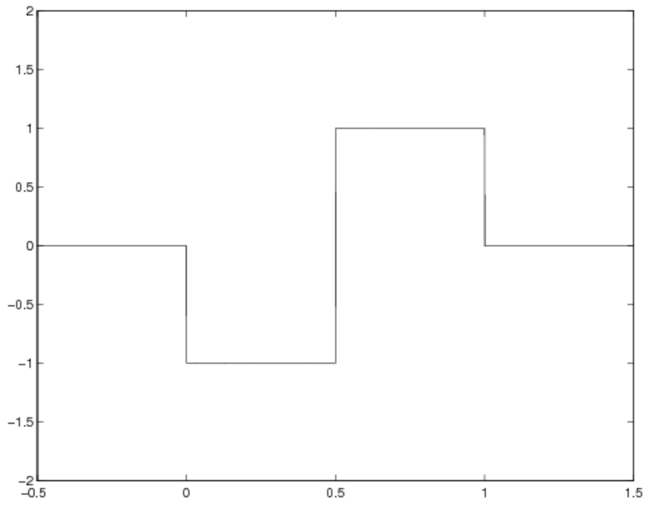
Equation:

$$\psi(x)^{\text{Haar}} = 2^{-1/2}\left(\phi^{\text{Haar}}(2x-1) - \phi^{\text{Haar}}(2x)\right) = 1_{\left[\frac{1}{2},1\right)}(x) - 1_{\left[0,\frac{1}{2}\right)}(x)\ .$$

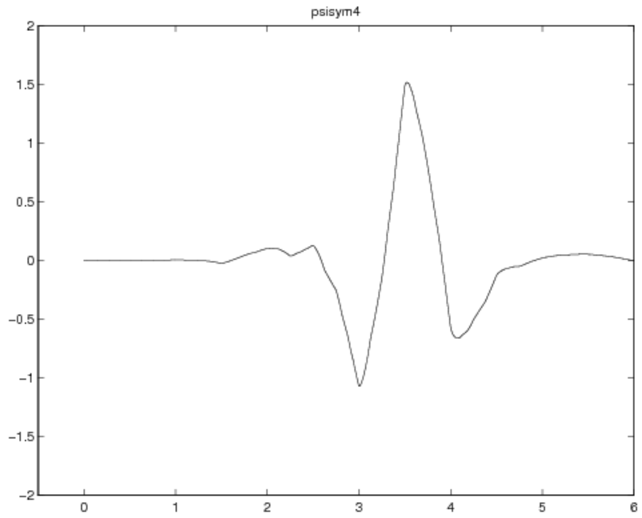
The Haar wavelet has only one vanishing moment and consequently is optimal only to represent functions having a low degree of regularity, like, for example, $\beta-$Hölder functions with $0 < \beta < 1$.

Daubechies constructed in [link], [link] compactly supported wavelets which have more than one vanishing moment. Compactly supported wavelets are desirable from a numerical point of view, while having more than one vanishing moment allows to reconstruct exactly polynomials of higher order. These wavelets cannot, in general, be written in a closed analytic form. However, their graph can be computed with arbitrarily high precision using a subdivision scheme algorithm. [link](b) represents the Daubechies Least Asymmetric wavelet with $N = 4$ vanishing moments.

Some orthogonal basis functions: (a) the Haar wavelet function bases with $N = 1$ vanishing moments, (b) the Least Asymmetric wavelet function of Daubechies [link], [link], with $N = 4$ vanishing moments.



(a) $N = 1$



(b)$N = 4$

This figure also illustrates the reason behind the name $wave \leq t\prime$: $\sin cewave \leq tsarefunctionswithacerta \in \nu mberofvanish \in gmoments, theyhavetheshapeo$ little wave' or `wavelet'.

Biorthogonal bases

Having an orthogonal MRA puts strong constraints on the construction of a wavelet basis. For example, the Haar wavelet is the only real-valued function which is compactly supported and symmetric. However, if we relax orthogonality for **biorthogonality**, then it becomes possible to have real-valued wavelet bases of fixed but arbitrary **high order** (see Definition 1 from Approximation of Functions) which are symmetric and compactly supported [link]. In a biorthogonal setting, a dual scaling function $\widetilde{\phi}$ and a dual wavelet function $\widetilde{\psi}$ exist. They generate a dual MRA with subspaces $\widetilde{V}_j$ and complement spaces $\widetilde{W}_j$ such that

Equation:

$$\widetilde{V}_j \perp W_j \quad \text{and} \quad V_j \perp \widetilde{W}_j \ .$$

In other words,

Equation:

$$\left\langle \widetilde{\phi}, \psi(\cdot - k) \right\rangle = 0 \quad \text{and} \quad \left\langle \phi, \widetilde{\psi}(\cdot - k) \right\rangle = 0$$

Moreover, the dual functions also have to satisfy

Equation:

$$\left\langle \widetilde{\phi}, \phi(\cdot - k) \right\rangle = \delta_{k,0} \quad \text{and} \quad \left\langle \widetilde{\psi}, \psi(\cdot - k) \right\rangle = \delta_{k,0} \ ,$$

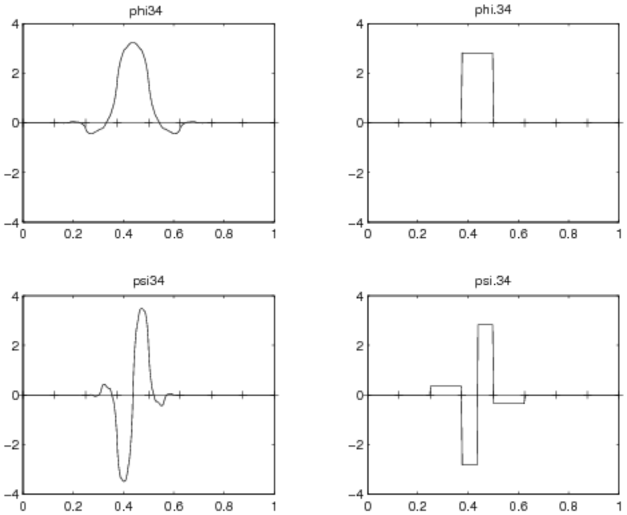
where $\delta_{k,0}$ is the Kronecker symbol. By construction, the dual scaling and wavelet functions satisfy a refinement equation, similarly to the equations [link] and [link].

In this work, we use the following convention: the dual MSD will be used to decompose a function (or a signal), while the original, or primal MSD reconstructs the function. This yields the following representation of a function $f \in L_2(\mathbb{R})$

Equation:

$$f(x) = \sum_k \left\langle f, \widetilde{\phi}_{j_0,k} \right\rangle \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_k \left\langle f, \widetilde{\psi}_{jk} \right\rangle \psi_{jk}(x) \ .$$

[link] shows an example of a biorthogonal wavelet basis built by Cohen, Daubechies and Feauveau in [link], (called CDF-wavelets hereafter).



Primal and dual scaling and wavelet functions for the (3,1)-Cohen-Daubechies-Feauveau (CDF) biorthogonal basis. The primal wavelet function ψ has one vanishing moment while the dual wavelet $\widetilde{\psi}$ has three vanishing moments.

One-dimensional wavelet transform

Suppose we are given as signal the projection of a function onto the space V_{j+1}:

Equation:

$$\mathscr{P}_{j+1} f = \sum_k s_{j+1,k} \phi_{j+1,k}(x), \quad s_{j+1,k} = \left\langle f\,, \widetilde{\phi}_{j+1,k} \right\rangle \,.$$

Using the dual refinement equations, we have:

Equation:

$$\begin{aligned} s_{j,k} = \left\langle f\,, \widetilde{\phi}_{j,k} \right\rangle &= \left\langle f\,, \sum_l \widetilde{h}_l \phi_{j+1,2k+l} \right\rangle \\ &= \sum_k \widetilde{h}_{l-2k} s_{j+1,l} \;, \end{aligned}$$

where the coefficients s_{jk} are called **scaling coefficients**, since they are related to scaling functions. Similarly, the **wavelet** or **detail coefficients**d_{jk} are obtained as

Equation:

$$d_{jk} = \left\langle f\,, \widetilde{\psi}_{jk} \right\rangle = \sum_k \tilde{g}_{l-2k} s_{j+1,l} \;.$$

The coefficients s_{jk} and d_{jk} are obtained from $s_{j+1,l}$ by `moving average' schemes, using the filter coefficients $\left\{ \widetilde{h}_l \right\}$ and $\{\tilde{g}_l\}$ as `weights', with the exception that these moving averages are sampled only at the even integers, i.e. a downsampling is performed. Such transform allows, once we have computed $s_{J,k}$= $\left\langle f, \widetilde{\phi}_{J,k} \right\rangle$ for a fine level $J \in \mathbb{N}$, to compute s_{jk} and d_{jk} for all coarser levels $j < J$ without evaluating the integrals.

Suppose now we are given the values of f at $n = 2^J$ equispaced design points. The scaling functions $\widetilde{\phi}_{J,k}$, $k = 0, ..., 2^J - 1$, are compactly supported and localized around $2^{-J}k$. Hence the coefficients $\left\langle f, \widetilde{\phi}_{J,k} \right\rangle$ are weighted and scaled average of f on a neighborhood of $2^{-J}k$ which becomes smaller as J tends to infinity. Consequently, it makes sense to replace the integral $\left\langle f, \widetilde{\phi}_{J,k} \right\rangle$ by the (scaled) value of f at $2^{-J}k$. More complicate quadrature formulae have been developed in [link], [link], [link].

With $s_j := \left\{ s_{jk}; k = 0, ..., 2^j - 1 \right\}$ and $d_j := \left\{ d_{jk}; k = 0, ..., 2^j - 1 \right\}$, the forward (or analyzing) wavelet transform given by [link]-[link] can be rewritten as

Equation:

$$s_j = \widetilde{H}_j^* s_{j+1} \quad \text{and} \quad d_j = \widetilde{G}_j^* s_{j+1} \ ,$$

where $\widetilde{H}_j^*$ denotes the Hermitian conjugate of $\widetilde{H}_j$.

The inverse (or synthesis) transform is found by using the primal refinement equations and the fact that $V_{j+1} = V_j \oplus W_j$.

Equation:

$$\begin{aligned} \mathscr{P}_{j+1} f &= \sum_l s_{j+1,l} \phi_{j+1,l} = \sum_k s_{j,k} \phi_{j,k} + \sum_k d_{j,k} \psi_{j,k} \\ &= \sum_k s_{j,k} \sum_l h_l \phi_{j+1,2k+l} + \sum_k d_{j,k} \sum_l g_l \phi_{j+1,2k+l} \\ &= \sum_l \phi_{j+1,l} \left(\sum_k h_{l-2k} s_{j,k} + \sum_k g_{l-2k} d_{jk} \right) \ , \end{aligned}$$

from which it follows that
Equation:

$$s_{j+1,l} = \sum_k h_{l-2k} s_{jk} + \sum_k g_{l-2k} d_{jk} \ .$$

In matrix form, we have
Equation:

$$s_{j+1} = H_j s_j + G_j d_j \ .$$

In the finite and classical setting, the matrices H_j, G_j, $\widetilde{H}_j$ and $\widetilde{G}_j$ are of size $2^{j+1} \times 2^j$. Moreover, if the basis functions are compactly supported, the four filters (h_l, g_l, $\widetilde{h}_l$, $\tilde{g}_l$) have only a finite number of nonzero elements, and hence all these matrices are banded.

Example: Haar wavelet transform

In case of the orthogonal Haar transform, $\widetilde{H}_j^* = H_j^*$ and is of the form
Equation:

$$\widetilde{H}_j^* = \begin{bmatrix} h_0 & h_1 & & & & & \\ & & h_0 & h_1 & & & \\ & & & & \dots & & \\ & & & & & h_0 & h_1 \end{bmatrix}$$

since only h_0 and h_1 are different from zero : $h_0 = h_1 = 1/\sqrt{2}$. The high-pass filter $\{g_l\}$ is such that $g_0 = -1/\sqrt{2}$ and $g_1 = 1/\sqrt{2}$. The forward transform [link]-[link] reduces to
Equation:

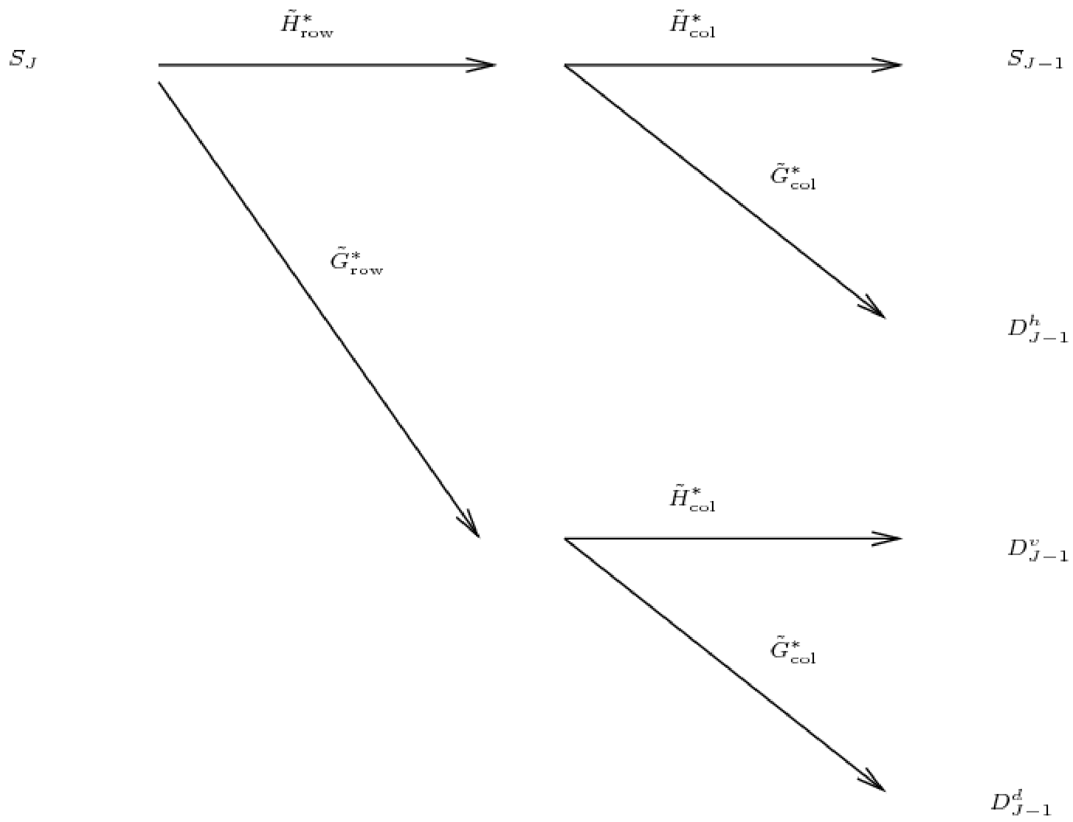
$$\begin{aligned} s_{j,k} &= \frac{1}{\sqrt{2}} s_{j+1,2k+1} + \frac{1}{\sqrt{2}} s_{j+1,2k} \\ d_{j,k} &= \frac{1}{\sqrt{2}} s_{j+1,2k+1} - \frac{1}{\sqrt{2}} s_{j+1,2k} \ , \end{aligned}$$

and the reconstruction is given by
Equation:

$$\begin{aligned} s_{j+1,2k} &= \frac{1}{\sqrt{2}} s_{j,k} - \frac{1}{\sqrt{2}} d_{j,k} \\ s_{j+1,2k+1} &= \frac{1}{\sqrt{2}} s_{j,k} + \frac{1}{\sqrt{2}} d_{j,k} \ . \end{aligned}$$

Two-dimensional wavelet transform

The wavelet transform has been successfully applied to compress images, which are modelled as functions defined on a regular two-dimensional grid.


S_J
$\tilde{H}^*_{\mathrm{row}}$
$\tilde{H}^*_{\mathrm{col}}$
S_{J-1}
$\tilde{G}^*_{\mathrm{col}}$
$\tilde{G}^*_{\mathrm{row}}$
D^h_{J-1}
$\tilde{H}^*_{\mathrm{col}}$
D^v_{J-1}
$\tilde{G}^*_{\mathrm{col}}$
D^d_{J-1}

Two-dimensional wavelet transform: first the filters are applied on the column of the matrix S_J, this produces two matrices. The filters are applied a second time on the columns of these two matrices, resulting in four elements: a matrix of scaling coefficients, and three detail matrices.

The easiest way to build a two-dimensional MRA is probably to use tensor products of spaces, see [link], [link]. In terms of wavelet transforms, this leads to applying two times a one-dimensional transform: first on the `row' of the signal matrix $S_J$, and second on the `columns' of the resulting two matrices, see [link]. In this figure, we see that, at each level of the decomposition, three types of detail coefficients are produced: D_j^h, D_j^v and D_j^d. These superscripts recall that, in an image, **horizontal** edges will lead to large values of D_j^h, **vertical** edges will show up in D_j^v and D_j^d will be sensitive to **diagonal** lines.

However, such a transform is not able to compress efficiently an image that contains curves. More complex bidimensional bases are now proposed in the literature to better model discontinuities along curves, see for example [link], [link], [link], [link].

Non-decimated wavelet transform

Suppose we have some signal $\{y_i\}$ observed at some equispaced design points : $y_i = f\left(i/n\right), i = 1, ..., n$ with $n = 2^J,\ J \in \mathbb{N}$. The transform presented in the previous section is sometimes called `decimated' because, for each scale j, the coefficients d_{jk} give only some information about the signal near the positions $x = 2^{-j}k$, and not near all the existing design points $2^{-J}k = k/n$.

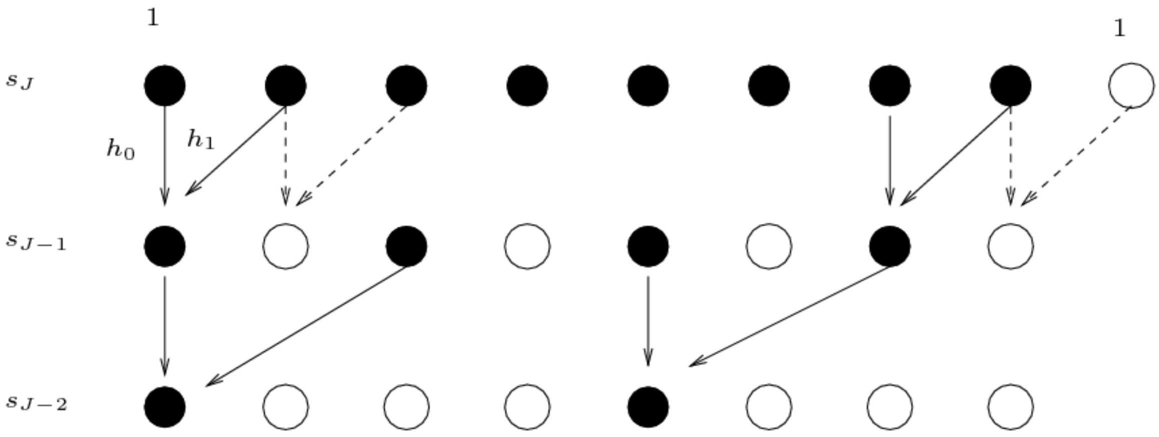
For this reason, the decimated wavelet transform lacks the property of translation invariance: given $t_0 \in \mathbb{R}$, the wavelet decomposition of $f(.\,)$ and of $f(.\,-t_0)$ are in general completely different. This may lead to some unwanted pseudo-Gibbs oscillations near a discontinuity which is not localized at a dyadic point [link].

One remedy to this drawback consists in using a non-decimated wavelet transform (NDWT) , also called **translation-invariant (TI)** [link] or **stationary** [link]. The idea behind the NDWT is to a perform a discrete wavelet transform, not only of the original sequence $\{y_i\}_{i=1}^n$, but of all the possible shifted sequences $\left(S_h y\right)_t = y_{(t+h) \bmod n}$. In terms of wavelet functions, this transform corresponds to a set of functions

Equation:

$$\widetilde{\psi}_{jk}\left(x\right) = \widetilde{\psi}\left(2^j\left(x - 2^{-J}k\right)\right), \quad j = j_0, ..., J-1,\ k = 0, ..., 2^J - 1\ .$$

At a given scale j, the NDWT coefficients are thus present at all the locations k/n for $k = 1, ..., n$ and give information about the signal at each observed design point. In other words, the non-decimated transform fills in the gap introduced in the decimated transform, see [link].



Schema illustrating the translation-invariant version of the Haar transform. The points marked by ● are the one computed for the decimated Haar transform. At level J, one circulant shift is performed: the first observation is put at the end of the observed signal, and a second decimated transform is performed on the shifted data (yielding the points marked by ○ at level $J-1$). This process is iterated at the coarser levels, producing detail coefficients at all the points.

Since we have $J - j_0$ scales and at each scales n detail coefficients, the NDWT gives an overdetermined representation of the original signal $\{y_i\}_{i=1}^{n}$ and the wavelet coefficients $\{d_{jk}, j = 0, ..., J-1, k = 1, ..., n\}$ are related to many different bases. Therefore the inverse operator will not be unique. A particular inverse, the average basis, corresponds to systematically average out the inverse wavelet transform obtained from each decimated wavelet transform that constitutes the translation-invariant transform. This makes the reconstruction robust with respect to a bad choice of a particular basis. Moreover, this average basis provides a smoother reconstruction than the original, decimated, transform [link], [link].

It allows for a (nearly) exact reconstruction of piecewise linear functions, instead of piecewise constant functions for the decimated Haar transform [link].

Approximation of Functions

We first give a definition of the order of a multiresolution analysis.

Note:(Order of a MRA in the classical setting) A multiresolution analysis is said to be of order $\widetilde{N}$ if the primal scaling function ϕ reproduces polynomials up to degree $\widetilde{N}-1$, i.e., $\text{For} \quad 0 \le p < \widetilde{N},\ \exists c_k \in \mathbb{R} \text{ such that } \ x^p = \sum_k c_k \phi\left(x-k\right)\ .$

The associated dual wavelet $\widetilde{\psi}$ has then $\widetilde{N}$ vanishing moments. In the classical setting, it is proved that the order of a MRA and the regularity of the scaling function are linked: the larger $\widetilde{N}$, the higher the regularity of ϕ. Symmetrically to [link], the order of the dual MRA is N if $\widetilde{\phi}$ can reproduce polynomials up to degree $N-1$. Figure 2 from Multiresolution analysis and wavelets shows an example of a biorthogonal basis where $\widetilde{N}=3$ and $N=1$. It illustrates the link between a high number of vanishing moments of the dual wavelet $\widetilde{\psi}$ and the regularity of the primal scaling function ϕ.

The main objective when decomposing a function in a wavelet series is to create a sparse representation of the function, that is, to obtain a decomposition where only a few number of detail coefficients are $lar \geq\prime, whi \leq themaj \text{ or } ityofthecoefficientsareclose \rightarrow zer \odot By$large', we mean that the absolute value of the detail coefficient is large.

Near a singularity, large detail coefficients at different levels will be needed to recover the discontinuity. However, between points of singularity, we can hope to have small detail coefficients, in particular if the analyzing wavelets $\widetilde{\psi}_{jk}$ have a large number $\widetilde{N}$ of vanishing moments. Indeed, suppose the function f to be decomposed is analytic on the interval I without discontinuity. Since $\left\langle x^p\, , \widetilde{\psi}_{jk} \right\rangle = 0$ for $p = 0, \ldots, \widetilde{N}-1$, we are sure that the first $\widetilde{N}$ terms of a Taylor expansion of f will not give a contribution to the wavelet coefficient $\left\langle f\, , \widetilde{\psi}_{jk} \right\rangle$ provided that the support of $\widetilde{\psi}_{jk}$ does not contain any singularities of the function f.

This sparse representation explains why classical wavelets provide smoothness characterization of function spaces like the Hölder and Sobolev spaces [link], but also of more general Besov spaces, which may contain functions of inhomogeneous regularity [link], [link], [link], [link], [link].

We illustrate this characterization property with the case of $\beta-$Hölder functions.

Definition 2
The class $\Lambda^{\beta}\left(L\right)$ of Hölder continuous functions is defined as follows:

1. if $\beta \leq 1, \Lambda^{\beta}\left(L\right) = \left\{f : |f(x) - f(y)| \leq L|x-y|^{\beta}\right\}$
2. if $\beta > 1, \Lambda^{\beta}\left(L\right) = \left\{f : \left|f^{(\lfloor\beta\rfloor)}\left(x\right) - f^{(\lfloor\beta\rfloor)}\left(y\right)\right| \leq L|x-y|^{\beta'} ; \left|f^{(\lfloor\beta\rfloor)}\left(x\right)\right| \leq M\right\}$, where $\lfloor\beta\rfloor$ is the largest integer less than β and $\beta' = \beta - \lfloor\beta\rfloor$.

The global Hölder regularity of a function can be characterized as follows [link], [link].

Note: Let $f \in \Lambda^{\beta}\left(L\right)$, and suppose that the (orthogonal) wavelet function ψ has r continuous derivatives and r vanishing moments with $r > \beta$. Then

Equation:

$$|\langle f\, , \psi_{jk}\rangle| \leq C2^{-j(\beta+1/2)}\ .$$

A similar characterization exists for continuous and Sobolev functions [link], [link].

In the orthogonal setting, the wavelet ψ must be regular **and** have a high number of vanishing moments. On the contrary, in the biorthogonal expansion equation 5 from Multiresolution analysis and wavelets, it is mostly of interest to have a dual wavelet $\widetilde{\psi}$ with a high number of vanishing moments, and hence a regular primal scaling and wavelet functions. On the primal side, it is sufficient to have only one vanishing moment for wavelet denoising, and consequently $\widetilde{\psi}$ may not be very regular. In this case, the wavelet coefficient $\left\langle f\, , \widetilde{\psi}_{jk}\right\rangle$ with the less regular wavelet $\widetilde{\psi}_{jk}$ can be used to characterize $f \in \Lambda^{\beta}\left(L\right)$ with $0 < \beta < \widetilde{N}$, even if $\beta > N = 1$: with a biorthogonal basis, regular functions can be characterized by their inner products with much less regular functions.

Nonparametric regression with wavelets

In this section, we consider only real-valued wavelet functions that form an orthogonal basis, hence $\phi \equiv \widetilde{\phi}$ and $\psi \equiv \widetilde{\psi}$. We saw in Orthogonal Bases from Multiresolution analysis and wavelets how a given function belonging to $L_2\left(\mathbb{R}\right)$ could be represented as a wavelet series. Here, we explain how to use a wavelet basis to construct a nonparametric estimator for the regression function m in the model

Equation:

$$Y_i = m\left(x_i\right) + \epsilon_i,\ i = 1, ..., n,\ n = 2^J,\ J \in \mathbb{N}\ ,$$

where $x_i = \frac{i}{n}$ are equispaced design points and the errors are i.i.d. Gaussian, $\epsilon_i\ \sim\ N\left(0, \sigma_\epsilon^2\right)$.

A wavelet estimator can be **linear** or **nonlinear**. The linear wavelet estimator proceeds by projecting the data onto a coarse level space. This estimator is of a kernel-type, see "Linear smoothing with wavelets". Another possibility for estimating m is to detect which detail coefficients convey the important information about the function m and to put equal to zero all the other coefficients. This yields a nonlinear wavelet estimator as described in "Nonlinear smoothing with wavelets".

Linear smoothing with wavelets

Suppose we are given data $\left(x_i, Y_i\right)_{i=1}^n$ coming from the model [link] and an orthogonal wavelet basis generated by $\{\phi, \psi\}$. The linear wavelet estimator proceeds by choosing a cutting level j_1 and represents an estimation of the projection of m onto the space V_{j_1}:

Equation:

$$\widehat{m}\left(x\right) = \sum_{k=0}^{2^{j_0}-1} \hat{s}_{j_0,k}\phi_{j_0,k}\left(x\right) + \sum_{j=j_0}^{j_1-1}\sum_{k=0}^{2^j-1} \widehat{d}_{j,k}\psi_{j,k}\left(x\right) = \sum_k \hat{s}_{j_1,k}\phi_{j_1,k}\left(x\right),$$

with j_0 the coarsest level in the decomposition, and where the so-called **empirical coefficients** are computed as
Equation:

$$\hat{s}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} Y_i \ \phi_{jk}(x_i) \quad \text{and} \quad \widehat{d}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} Y_i \ \psi_{jk}(x_i) \ .$$

The cutting level j_1 plays the role of a smoothing parameter: a small value of j_1 means that many detail coefficients are left out, and this may lead to oversmoothing. On the other hand, if j_1 is too large, too many coefficients will be kept, and some artificial bumps will probably remain in the estimation of $m(x)$.

To see that the estimator [link] is of a kernel-type, consider first the projection of m onto V_{j_1}:
Equation:

$$\begin{aligned} \mathscr{P}_{V_{j_1}} m(x) &= \sum_k \left(\int m(y) \phi_{j_1,k}(y) dy \right) \phi_{j_1,k}(x) \\ &= \int K_{j_1}(x,y) m(y) dy \ , \end{aligned}$$

where the (convolution) kernel $K_{j_1}(x,y)$ is given by
Equation:

$$K_{j_1}(x,y) = \sum_k \phi_{j_1,k}(y) \phi_{j_1,k}(x) \ .$$

Härdle **et al.** [link] studied the approximation properties of this projection operator. In order to estimate [link], Antoniadis **et al.** [link] proposed to take:
Equation:

$$\begin{aligned}\widehat{\mathscr{P}_{V_{j_1}}} m(x) &= \sum_{i=1}^{n} Y_i \int_{(i-1)/n}^{i/n} K_{j_1}(x,y)dy \\ &= \sum_{k} \sum_{i=1}^{n} Y_i \left(\int_{(i-1)/n}^{i/n} \phi_{j_1,k}(y)dy \right) \phi_{j_1,k}(x) \ .\end{aligned}$$

Approximating the last integral by $\frac{1}{n}\phi_{j_1,k}(x_i)$, we find back the estimator $\widehat{m}(x)$ in [link].

By orthogonality of the wavelet transform and Parseval's equality, the L_2-risk (or integrated mean square error IMSE) of a linear wavelet estimator is equal to the l_2-risk of its wavelet coefficients:

Equation:

$$\begin{aligned}\text{IMSE} = E\|\widehat{m} - m\|_{L_2}^2 &= \sum_k E\left[\hat{s}_{j_0,k} - s^{\circ}_{j_0,k}\right]^2 + \sum_{j=j_0}^{j_1-1} \sum_k E\left[\widehat{d}_{jk} - d^{\circ}_{jk}\right]^2 \\ &+ \sum_{j=j_1}^{\infty} \sum_k {d^{\circ}_{jk}}^2 = S_1 + S_2 + S_3 \ ,\end{aligned}$$

where

Equation:

$$s^{\circ}_{jk} := \langle m \, , \phi_{jk} \rangle \quad \text{and} \quad d^{\circ}_{jk} = \langle m \, , \psi_{jk} \rangle$$

are called `theoretical' coefficients in the regression context. The term $S_1 + S_2$ in [link] constitutes the stochastic bias whereas S_3 is the deterministic bias. The optimal cutting level is such that these two bias are of the same order. If m is $\beta-$Hölder continuous, it is easy to see that the optimal cutting level is $j_1(n) = O\left(n^{1/(1+2\beta)}\right)$. The resulting optimal IMSE is of order $n^{-\frac{2\beta}{2\beta+1}}$. In practice, cross-validation methods are often used to determine the optimal level j_1 [link], [link].

Nonlinear smoothing with wavelets

Hard-, soft-thresholding and wavelet estimator

Given the regression model [link], we can decompose the empirical detail coefficient $\widehat{d}_{jk}$ in [link] as

Equation:

$$\begin{aligned}\widehat{d}_{jk} &= \frac{1}{n}\sum_{i=1}^{n} m\left(x_i\right)\psi_{jk}\left(x_i\right) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i\psi_{jk}\left(x_i\right) \\ &= d_{jk} + \rho_{jk}\end{aligned}$$

If the function $m(x)$ allows for a sparse wavelet representation, only a few number of detail coefficients d_{jk} contribute to the signal and are non-negligible. However, every empirical coefficient $\widehat{d}_{jk}$ has a non-zero contribution coming from the noise part ρ_{jk}.

Note: Note the link between the coefficients d_{jk} in [link] and the theoretical coefficients d_{jk}° in [link]:

Equation:

$$\begin{aligned}d_{jk} &= \frac{1}{n}\sum_{i=1}^{n} m\left(x_i\right)\psi_{j,k}\left(x_i\right) \\ &= \int m\left(x\right)\psi_{jk}\left(x\right)dx + O\left(\frac{1}{n}\right) = d_{jk}^{\circ} + O\left(\frac{1}{n}\right).\end{aligned}$$

In words, d_{jk} constitutes a first order approximation (using the trapezoidal rule) of the integral d_{jk}°. For the scaling coefficients s_{jk}°, it can be proved

[link] that the order of accuracy of the trapezoidal rule is equal to $N-1$, where N is the order of the MRA associated to the scaling function.

Suppose the noise level is not too high, so that the signal can be distinguished from the noise. Then, from the sparsity property of the wavelet, only the largest detail coefficients should be included in the wavelet estimator. Hence, when estimating an unknown function, it makes sense to include only those coefficients that are larger than some specified threshold value t:

Equation:

$$\eta_H\left(\widehat{d}_{jk},t\right)=\widehat{d}_{jk}1_{\left\{\left|\widehat{d}_{jk}\right|>t\right\}}\;.$$

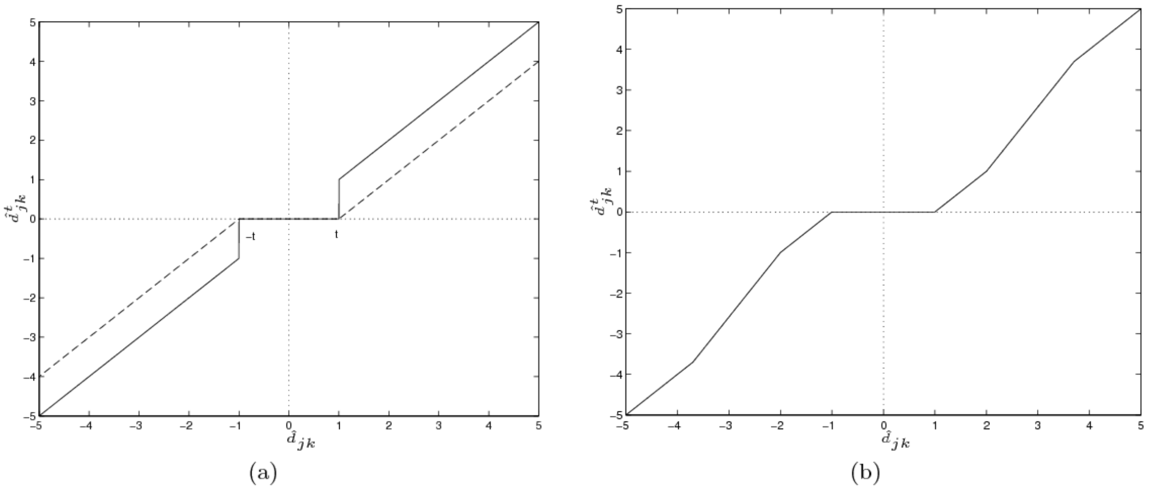
This `keep-or-kill' operation is called **hard thresholding**, see [link](a).

Now, since each empirical coefficient consists of both a signal part and a noise part, it may be desirable to shrink even the coefficients that are larger than the threshold:

Equation:

$$\widehat{d}_{jk}^{t}:=\eta_S\left(\widehat{d}_{jk},t\right)=\mathrm{sign}\left(\widehat{d}_{jk}\right)\left(\left|\widehat{d}_{jk}\right|-t\right)_+\;.$$

Since the function η_S is continuous in its first argument, this procedure is called **soft thresholding**. More complex thresholding schemes have been proposed in the literature [link], [link], [link]. They often appear as a compromise between soft and hard thresholding, see [link](b) for an example.



In (a) the hard thresholding is represented in plain line: a coefficient $\widehat{d}_{jk}$ with an absolute value below t is put equal to zero. The soft thresholding is given in dashed line: there coefficients with absolute value above the threshold t are shrunk of an amount equal to t. In (b), a more complex thresholding procedure, the SCAD threshold devised in Antoniadis and Fan [link] is represented.

For a given threshold value t and a thresholding scheme $\eta_{(.)}$, the nonlinear wavelet estimator is given by

Equation:

$$\widehat{m}\left(x\right)=\sum_{k}\hat{s}_{j_0k}\ \phi_{j_0k}\left(x\right)\ +\ \sum_{j,k}\eta_{(.)}\left(\widehat{d}_{jk},t\right)\ \psi_{jk}\left(x\right)\ ,$$

where j_0 denotes the **primary resolution level**. It indicates the level above which the detail coefficients are being manipulated.

Let now $\widehat{d}_{j}=\left\{\widehat{d}_{jk},k=0,...,2^{j}-1\right\}$ denote the vector of empirical detail coefficients at level j and similarly define $\hat{s}_{j}$. In practice a nonlinear wavelet estimator is obtained in three steps.

1. Apply the analyzing (forward) wavelet transform on the observations $\{Y_i\}_{i=1}^n$, yielding $\hat{s}_{j_0}$ and $\widehat{d}_j$, for $j = j_0, ..., J-1$.
2. Manipulate the detail coefficients above the level j_0, e.g. by soft-thresholding them.
3. Invert the wavelet transform and produce an estimation of m at the design points: $\left\{\widehat{m}\left(x_i\right)\right\}_{i=1}^n$.

If necessary, a continuous estimator $\widehat{m}$ can then be constructed by an appropriate interpolation of the estimated $\widehat{m}\left(x_i\right)$ values [link].

The choice of the primary resolution level in nonlinear wavelet estimation has the same importance as the choice of a particular kernel in local polynomial estimation, i.e., it is not the most important factor. It is common practice to take $j_0 = 2$ or $j_0 = 3$, although a cross-validation determination is of course possible [link].

The selection of a threshold value is much more crucial. If it is chosen too large, the thresholding operation will kill too many coefficients. Too few coefficients will then be included in the reconstruction, resulting in an oversmoothed estimator. Conversely, a small threshold value will allow many coefficients to be included in the reconstruction, giving a rough, or undersmoothed estimator. A proper choice of the threshold involves thus a careful balance between smoothness and closeness of fit.

In case of an orthogonal transform and i.i.d. white noise, the same threshold can be applied to all detail coefficients, since the errors in the wavelet domain are still i.i.d. white noise. However, if the errors are stationary but correlated, or if the transform is biorthogonal, a level-dependent threshold is necessary to obtain optimal results [link], [link]. Finally, in the irregular setting, a level and location dependent threshold must be utilized.

Many efforts have been devoted to propose methods for selecting the threshold. We now review some of the procedures encountered in the literature.

Choice of the threshold

Universal threshold

The most simple method to find a threshold whose value is supported by some statistical arguments, is probably to use the so-called `universal threshold' [link], [link]

Equation:

$$t_{\mathrm{univ}} = \sigma_d \sqrt{2\,\log\,n}\ ,$$

where the only quantity to be estimated is σ_d^2, which constitutes the variance of the empirical wavelet coefficients. In case of an orthogonal transform, $\sigma_d = \sigma_\epsilon / \sqrt{n}$.

In a wavelet transform, the detail coefficients at fine scales are, with a small fraction of exception, essentially pure noise. This is the reason why Donoho and Johnstone proposed in [link] to estimate σ_d in a robust way using the median absolute deviation from the median (MAD) of $\widehat{d}_{J-1}$:

Equation:

$$\widehat{\sigma}_d = \frac{\mathrm{median}\left(\left|\widehat{d}_{J-1} - \mathrm{median}\left(\widehat{d}_{J-1}\right)\right|\right)}{0.6745}\ .$$

If the universal threshold is used in conjunction with soft thresholding, the resulting estimator possesses a noise-free property: with a high probability, an appropriate interpolation of $\left\{\widehat{m}\left(x_i\right)\right\}$ produces an estimator which is at least as smooth as the function m, see Theorem 1.1 in [link]. Hence the reconstruction is of good visual quality, so that Donoho and Johnstone called the procedure `VisuShrink' [link]. Although simple, this estimator enjoys a near-minimax adaptivity property, see "Adaptivity of wavelet estimator". However, this near-optimality is an asymptotic one: for small sample size t_{univ} may be too large, leading to a poor mean square error.

Oracle inequality

Consider the soft-thresholded detail coefficients $\widehat{d}^t$. Another approach to find an optimal threshold is to look at the l_2-risk

Equation:

$$\mathscr{R}\left(\widehat{d}^t, d\right) = E\sum_{(j,k)}\left(\widehat{d}^t_{jk} - d_{jk}\right)^2 = E\|\widehat{d}^t - d\|^2_{l_2},$$

and to relate this risk with the one of an ideal risk $\mathscr{R}_{\text{ideal}}$. The ideal risk is the risk obtained if an oracle tells us exactly which coefficients to keep or to kill.

In [link], Donoho and Johnstone showed that, when using the universal threshold, the following oracle inequality prevails

Equation:

$$\mathscr{R}\left(\widehat{d}^t, d\right) \le (2\,\log\,n + 1)\left(\frac{\sigma_\epsilon^2}{n} + \mathscr{R}_{\text{ideal}}\right).$$

However, this inequality is not optimal. Donoho and Johnstone looked for the optimal threshold $t^*(n)$ which leads to the smallest possible constant Λ_n^* in place of $2\,\log\,n + 1$. Such a threshold does not exist in closed form, but can be approximated numerically. For small sample size, it is sensibly smaller than the universal threshold.

SureShrink procedure

Given the expression [link] for the l_2-risk, it is natural to look for a threshold that minimizes an estimation of this risk.

By minimizing Stein's unbiased estimate of the risk [link] and using a soft thresholding scheme, the resulting estimator, called `SureShrink', is adaptive over a wide range of function spaces including Hölder, Sobolev, and Besov spaces, see "Adaptivity of wavelet estimator". That is, without any a priori knowledge on the type or amount of regularity of the function, the SURE procedure nevertheless achieves the optimal rate of convergence that one could attain by knowing the regularity of the function.

Other thresholding procedures

We mention some of the other thresholding or shrinkage procedures proposed in the literature.

Instead of considering each coefficient individually, Cai **et al.** [link], [link] consider blocks of empirical wavelet coefficients in order to make simultaneous shrinkage decisions about all coefficients within a block.

Another fruitful idea is to use the Bayesian framework. There a prior distribution is imposed on the wavelet coefficients d_{jk}. This prior model is designed to capture the sparseness of the wavelet expansion. Next, the function is estimated by applying some Bayes rules on the resulting posterior distribution of the wavelet coefficients, see for example [link], [link], [link], [link].

Antoniadis and Fan [link] treat the problem of selecting the wavelet coefficients as a penalized least squares issue. Let W be the matrix of an orthogonal wavelet transform and $Y := \{Y_i\}_{i=1}^{n}$. The detail coefficients $d := \{d_{jk}\}$ which minimize
Equation:

$$\|WY - d\|_{l_2}^2 + \sum_{j,k} p_\lambda\left(|d_{jk}|\right)$$

are used to estimate the true wavelet coefficients. In equation [link], $p_\lambda\left(\cdot\right)$ is a penalty function which depends on the regularization parameter λ. The authors provide a general framework, where different penalty functions correspond to different type of thresholding procedures (like, e.g., the soft- and hard- thresholding) and obtain oracle inequalities for a large class of penalty functions.

Other methods include threshold selection by hypothesis testing [link], cross-validation [link], or generalized cross-validation [link], [link], which is used to estimated the l_2-risk of the empirical detail coefficients.

Linear versus nonlinear wavelet estimator

In order to differenciate the behaviours of a linear and of a nonlinear wavelet estimator, we consider the Sobolev class $W_q^s\left(C\right)$ defined as

Equation:

$$W_q^s\left(C\right) = \left\{f : \|f\|_q^q + \|\frac{d^s}{dx^s}f\left(x\right)\|_q^q \leq C^2\right\},$$

and that we denote V in short. Assume we know that m, the function to be estimated, belongs to V. In the next section, we will release this assumption. The L_p-risk of an arbitrary estimator T_n based on the sample data is defined as $E\|T_n - m\|_p^p$, $1 \leq p < \infty$, whereas the L_p-minimax risk is given by

Equation:

$$R_n\left(V, p\right) = \inf_{T_n} \sup_{m \in V} E\|T_n - m\|_p^p,$$

where the infimum is taken over all measurable estimators T_n of m. Similarly, we define the linear L_p-minimax risk as

Equation:

$$R_n^{\text{lin}}\left(V, p\right) = \inf_{T_n^{\text{lin}}} \sup_{m \in V} E\|T_n^{\text{lin}} - m\|_p^p,$$

where the infimum is now taken over all **linear** estimators T_n^{lin}. Obviously, $R_n^{\text{lin}}\left(V, p\right) \geq R_n\left(V, p\right)$. We first state some definitions.

Note: The sequences $\{a_n\}$ and $\{b_n\}$ are said to be **asymptotically equivalent** and are noted $a_n \sim b_n$ if the ratio a_n / b_n is bounded away from zero and ∞ as $n \to \infty$.

Note: The sequence a_n is called optimal rate of convergence , (or **minimax rate of convergence**) on the class V for the L_p-risk if $a_n \sim R_n(V,p)^{1/p}$. We say that an estimator m_n of m attains the optimal rate of convergence if $\sup_{m\in V} E\|m_n - m\|_p^p \sim R_n\left(V,p\right)$.

In order to fix the idea, we consider only the L_2-risk in the remaining part of this section, thus $p := 2$.

In [link], [link], the authors found that the optimal rate of convergence attainable by an estimator when the underlying function belongs to the Sobolev class W_q^s is $a_n = n^{\frac{-s}{2s+1}}$, hence $R_n\left(V,2\right) = n^{\frac{-2s}{2s+1}}$. We saw in "Linear smoothing with wavelets" that linear wavelet estimators attain the optimal rate for $s-$Hölder function in case of the L_2-risk (also called `IMSE'). For a Sobolev class W_q^s, the same result holds provided that $q \geq 2$. More precisely, we have the two following situations.

1. If $q \geq 2$, we are in the so-called **homogeneous** zone. In this zone of spatial homogeneity, linear estimators can attain the optimal rate of convergence $n^{-s/(2s+1)}$.
2. If $q < 2$, we are in the **non-homogeneous** zone, where linear estimators do not attain the optimal rate of convergence. Instead, we have:
   Equation:

$$R_n^{\text{lin}}\left(V,2\right)/R_n\left(V,2\right) \to \infty, \text{ as } n \to \infty.$$

The second result is due to the spatial variability of functions in Sobolev spaces with small index q. Linear estimators are based on the idea of spatial homogeneity of the function and hence do perform poorly in the presence of non-homogeneous functions. In contrast, even if $q < 2$, the SureShrink estimator attains the minimax rate [link]. The same type of results holds for more general Besov spaces, see for example [link], Chapter 10.

Adaptivity of wavelet estimator

We just saw that a nonlinear wavelet estimator is able to estimate in an optimal way functions of inhomogeneous regularity. However, it may not be sufficient to know that for m belonging to a given space, the estimator performs well. Indeed, in general we do not know which space the function belongs to. Hence it is of great interest to consider a **scale** of function classes and to look for an estimator that attains **simultaneously** the best rates of convergence across the whole scale. For example, the L_q-Sobolev scale is a set of Sobolev function classes $W_q^s\left(C\right)$ indexed by the parameters s and C, see [link] for the definition of a Sobolev class. We now formalize the notion of an adaptive estimator.

Let A be a given set and let $\{\mathcal{F}_\alpha, \alpha \in A\}$ be the scale of functional classes $\mathcal{F}_\alpha$ indexed by $\alpha \in A$. Denote $R_n\left(\alpha, p\right)$ the minimax risk over $\mathcal{F}_\alpha$ for the L_p-loss:

Equation:

$$R_n\left(\alpha, p\right) = \inf_{\widehat{m}_n} \sup_{m \in \mathcal{F}_\alpha} E\|\widehat{m}_n - m\|_p^p.$$

Note: The estimator $\overset{*}{m}_n$ is called **rate adaptive** for the L_p-loss and the scale of classes $\mathcal{F}_\alpha, \alpha \in A$ if for any $\alpha \in A$ there exists $c_\alpha > 0$ such that

Equation:

$$\sup_{m \in \mathcal{F}_\alpha} E\|\overset{*}{m}_n - m\|_p^p \leq c_\alpha R_n\left(\alpha, p\right) \forall n \geq 1.$$

The estimator $\overset{*}{m}_n$ is called **adaptive up to a logarithmic factor** for the L_p-loss and the scale of classes $\mathcal{F}_\alpha, \alpha \in A$ if for any $\alpha \in A$ there exist $c_\alpha > 0$ and $\gamma = \gamma_\alpha > 0$ such that

Equation:

$$\sup_{m \in \mathscr{F}_\alpha} E\|\overset{*}{m}_n - m\|_p^p \leq c_\alpha (\log n)^\gamma R_n\,(\alpha, p)\,\forall n \geq 1.$$

Thus, adaptive estimators have an optimal rate of convergence and behave as if they know in advance in which class the function to be estimated lies.

The VisuShrink procedure is adaptive up to a logarithmic factor for the L_2-loss over every Besov, Hölder and Sobolev class that is contained in $C[0,1]$, see Theorem 1.2 in [link]. The SureShrink estimator does better: it is adaptive for the L_2-loss, for a large scale of Besov, Hölder and Sobolev classes, see Theorem 1 in [link].

Conclusion

In this chapter, we saw the basic properties of standard wavelet theory and explained how these are related to the construction of wavelet regression estimators.